UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

September 21, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES B-10

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: James Farber JF
Sample Design Team

Subject: Accuracy and Coverage Evaluation Survey: Consistency of Post-Stratification Variables (Prototype)

The attached document is a prototype of the report that we will prepare, per your request, following completion of applicable Accuracy and Coverage Evaluation Survey (A.C.E.) operations. The completed report is intended to aid the Executive Steering Committee on A.C.E. Policy (ESCAP) in its recommendation regarding the release of the statistically corrected data or the data without statistical correction as the P.L. 94-171 data. This report, together with other reports, will assess the operations and results of both the initial Census and the A.C.E. Both sets of assessments will be available to the ESCAP to aid the Committee in reaching its recommendation regarding the use of the statistically corrected data.

The attached prototype contains empty table shells that will assess specific aspects of the applicable operations. This report focuses on the consistency of post-stratification variables between the P sample and E sample. The analysis is limited to P-sample and E-sample cases that matched following person matching.

It is important to note that the conduct of the operations may lead us to modify the attached format by including additional information. It is also likely that descriptions and definitions will be enhanced or the data items could undergo revision. Conversely, we may conclude, for a variety of reasons, that some of the information set forth in the attached prototype may not be available. The attached document sets forth our conclusions prior to completion of the A.C.E. about what information would properly inform the ESCAP on this subject, but is subject to modification.

Accuracy and Coverage Evaluation 2000: Consistency of Post-Stratification Variables

prepared by James Farber

Introduction

The Accuracy and Coverage Evaluation (A.C.E.) consists of two independent samples. The first is a sample of the population in selected A.C.E. sample areas, known as the Population or P sample. Matching these people to census records provides an estimate of the net proportion of the population missed in the census. The second is a sample of the census enumerations in the same A.C.E. sample areas, known as the Enumeration or E sample. Using the results of matching the P sample to the census, checking for duplication among the census records, and re-interviewing when needed to determine the correct inclusion of each E-sample record, an estimate of the net proportion of correctly enumerated records in the census can be determined.

The A.C.E. includes dual system estimates for up to 448 post-strata for the 50 states and the District of Columbia (Haines, 2000). Each P-sample person and E-sample person is assigned to a post-stratum. Ideally, a P-sample person who matches to an E-sample person will have consistent post-stratification variables: race, Hispanic origin, age, sex, and tenure. In reality, this may not occur. If a person does not have consistent characteristics in his P-sample and E-sample records, then that person could be in different post-strata when estimating the net proportions of people missed or correctly enumerated in the census.

The purpose of this report is to get an indication of the consistency of the post-stratification variables between the two systems. Persistent differences in the classification of persons in the census and the A.C.E. may introduce a potential bias into the coverage estimates. This bias is sometimes referred to as classification error.

One reason the two systems may differ is when a person has an unknown post-stratification variable that is filled in through characteristic imputation. Both the P sample and E sample are subject to characteristic imputation. Appendix 1 gives the criteria for determining if a characteristic has been imputed. There are separate criteria for the P sample and the E sample. For this analysis, if a characteristic is imputed for either the P sample or the E sample, then the case is considered imputed.

This report distinguishes between imputed and non-imputed characteristics. Some tables in this report display results only for total cases and for non-imputed cases. The corresponding results for imputed cases are the difference between the total and non-imputed. This decomposition clarifies the source of inconsistency. For imputed cases, inconsistency is usually attributable to the characteristic imputation procedure. For non-imputed cases, inconsistencies arise due to inconsistent reporting, which has many possible causes including the data collection mode, time lag from reference day, proxy responses, or data capture difficulties. This report does not explore these reasons for misclassification.

This report also excludes adjustments for P-sample matches to E-sample duplicates. In person matching, a P-sample case may possibly match to an E-sample case found to be a duplicate of another E-sample case or a non-E-sample census record. In this situation, only the matched P-sample and E-sample records will be used to assess consistency.

A similar analysis on consistency was conducted for the 1998 Census 2000 Dress Rehearsal and the 1995 Census Test (see Salganik, 1999 and Petroni, 1996A and 1996B). An analysis on the 1990 Post-Enumeration Survey (PES) data is desirable but information linking PES P-sample and E-sample records is not readily available, making such an analysis impossible at this time.

Measuring Consistency

A variable is defined as consistent when the information collected in the P and E samples is the same or results in the classification of the person to the same level of the post-stratification variable. For example, a person who reports her age as 28 in the P sample and 27 in the E sample would be classified in the 18 - 29 group of the age post-stratification variable, and thus that person's age is consistent even though it does not match exactly.

To measure the consistency of post-stratification variables, we are limited to looking at P-sample cases that matched to an E-sample case following the A.C.E. person matching operation. This means that P-sample cases that matched to a census enumeration not in the E sample are excluded from this analysis. Including such cases is technically feasible but difficult in practice, and the gain in assessing consistency would likely be minimal since there are relatively few of these cases compared to the number of matches. We will be able to detect any classification error problem using the large amount of readily available data from matched cases.

This report also looks at whether misclassifications are balanced. Inconsistencies that occur randomly and are balanced are of less concern than systematic switching from one group to another, an imbalanced scenario. Note, though, that even imbalanced inconsistency is a concern only when the matched person's two different post-strata have significantly different coverage properties. If the coverage correction factors for the two post-strata are very similar, then the misclassification has no practical effect on the A.C.E. population estimates. Classification error is a function of not only the amount of inconsistency but also the differences in coverage rates among the post-strata. This report does not include an analysis of the coverage rates of the post-strata in the various combinations of inconsistent cases.

This study may under-report the amount of inconsistency because the data include only matched cases. The non-matched people may be more inconsistent simply because they cannot be matched. Use caution when drawing conclusions about the entire population based on the consistency of only the matched people. Ideally, this report would include the non-matched cases to obtain an overall measure of inconsistency, but this is not possible. Assessing consistency requires that the P-sample and E-sample information be linked. Non-matched people do not have that link, and thus can not be studied.

Results

Post-Stratification Variables: The post-stratification variables considered in this analysis are tenure, age/sex, and race/Hispanic origin domain. All other post-stratification variables are geographically assigned variables that by definition are consistent between the P and E samples. A person is consistent if their P-sample and E-sample responses are in the same group of each post-stratification variable, as listed below:

Tenure

- Owner
- Non-owner

Age/Sex

- Under 18
- 18 - 29 Male
- 18 - 29 Female
- 30 - 49 Male
- 30 - 49 Female
- 50 + Male
- 50 + Female

Race/Hispanic Origin Domain: See Appendix 2 for more detail on these seven domains.

- Domain 1 American Indian or Alaska Native on reservations
- Domain 2 American Indian or Alaska Native off reservations
- Domain 3 Hispanic
- Domain 4 Non-Hispanic Black
- Domain 5 Native Hawaiian or Pacific Islander
- Domain 6 Non-Hispanic Asian
- Domain 7 Non-Hispanic White or "Some Other Race"

Table 1 below summarizes the consistency for each of these three post-stratification variables by imputation status. See Tables A-1, A-2, and A-3 in Attachment A for more detailed results. The total number of matched cases is the same for all three variables, but their distributions differ by imputation status.

These tables also show non-balanced inconsistent cases, which are the absolute difference of the inconsistent cases. For example, Table A-1 shows of the total matched cases there are

- ___ P-sample owners and E-sample non-owners and
- ___ P-sample non-owners and E-sample owners

Thus there are

- ___ inconsistent cases (____ percent of the total matches) and
- ___ non-balanced cases (____ percent of the total matches)

Table 1: Consistency of Matching P- and E-Sample Post-Stratification Variables

Variable	Total Cases	Consistent Cases	Inconsistent		Non-Balanced	
			Cases	Percent	Cases	Percent
Tenure						
Non-Imputed						
Imputed						
Age/Sex						
Non-Imputed						
Imputed						
Race/Origin Domain						
Non-Imputed						
Imputed						

448 Post-Strata: Tables B-1 through B-64 in Attachment B show consistency results for each of the 64 major post-stratum groups by the 7 age/sex groups.

References

Haines, D. (2000), "Accuracy and Coverage Evaluation Survey: Final Post-stratification Plan for Dual System Estimation," DSSD Census 2000 Procedures and Operations Memorandum Series Q.

Petroni, R. (1996A), "Disagreement of Characteristics Between R-Sample and Census Linked Cases for Oakland - Preliminary Findings," Internal Census Bureau Memorandum.

Petroni, R. (1996B), "Disagreement of Characteristics Between R-Sample and Census Linked Cases for Oakland - More Findings," Internal Census Bureau Memorandum.

Salganik, M. (1999), "Accuracy and Coverage Evaluation Survey: Consistency of Potential Poststratification Variables," DSSD Census 2000 Procedures and Operations Memorandum Series Q.

Table A-1: Consistency of Post-Stratification Variables: Tenure

Total Matched Cases		E Sample Owner	E Sample Non-Owner	Total	% Inconsistent
P Sample	Owner				
	Non-Owner				
	Total				
	% Inconsistent				
Non-Imputed Cases					
P Sample	Owner				
	Non-Owner				
	Total				
	% Inconsistent				

Table A-2: Consistency of Post-Stratification Variables: Age/Sex

Total Matched Cases		E Sample 0 - 17	18 - 29 M	18 - 29 F	30 - 49 M	30 - 49 F	50+ M	50 + F	Total	% Inconsistent
P Sample	0 - 17									
	18 - 29 M									
	18 - 29 F									
	30 - 49 M									
	30 - 49 F									
	50 + M									
	50 + F									
	Total									
	% Inconsistent									
Non-Imputed Matched Cases										
P Sample	0 - 17									
	18 - 29 M									
	18 - 29 F									
	30 - 49 M									
	30 - 49 F									
	50 + M									
	50 + F									
	Total									
	% Inconsistent									

Table A-3: Consistency of Post-Stratification Variables: Race/Hispanic Origin Domains

Total Matched Cases		E Sample							Total	% Incon.
		Domain 1	Domain 2	Domain 3	Domain 4	Domain 5	Domain 6	Domain 7		
P Sample	Domain 1									
	Domain 2									
	Domain 3									
	Domain 4									
	Domain 5									
	Domain 6									
	Domain 7									
	Total									
	% Inconsistent									

Table A-3: Consistency of Post-Stratification Variables: Race/Hispanic Origin Domains (cont.)

Non-Imputed Matched Cases		E Sample							Total	% Incon.
		Domain 1	Domain 2	Domain 3	Domain 4	Domain 5	Domain 6	Domain 7		
P Sample	Domain 1									
	Domain 2									
	Domain 3									
	Domain 4									
	Domain 5									
	Domain 6									
	Domain 7									
	Total									
	% Inconsistent									

Tables B - 1 through B - 64: Consistency of P and E Samples by Post-Stratum Group

Post-Stratum Definition	M & F 0 - 17	Males 18 - 29	Females 18 - 29	Males 30 - 49	Females 30 - 49	Males 50+	Females 50+	Total
Total P & E Sample Matches Total Cases Total Consistent Persons Proportion Consistent								
Total P & E Sample Matches - Non-Imputed Total Cases Total Consistent Persons Proportion Consistent								
Total P & E Sample Matches - Imputed Total cases Total Consistent Persons Proportion Consistent								

Characteristic Imputation

P Sample

For dual system estimation, P-sample records are imputed because of either an edit failure or a missing value Table 1.1 identifies what is considered an imputed value for the variables needed to assign P-sample records to the post-strata. The values in the table are the imputation flag values for these variables. The flags can be found on the Person Dual System Estimation P-sample Output Person File.

Table 1.1: Identifying Imputed Values for the P Sample

P-sample Characteristic	Reported Values	Imputed Values
Age	1 = No imputation	2 = Imputation because of edit failure 3 = Imputation because of missing value
Race	1 = No imputation	2 = Imputation because of edit failure 3 = Imputation because of missing value
Hispanic Origin	1 = No imputation	2 = Imputation because of edit failure 3 = Imputation because of missing value
Sex	1 = No imputation	2 = Imputation because of edit failure 3 = Imputation because of missing value
Tenure	1 = No imputation	2 = Imputation because of edit failure 3 = Imputation because of missing value

E Sample

For dual system estimation, E-sample records are imputed by one of two steps. First, we try to obtain the demographic and tenure variables from the Hundred Percent Census Edited File (HCEF) for each record. We identify if the HCEF has done any editing or imputation for these records. If the record does not match to the HCEF then the Missing Data process has a backup imputation system to impute missing values. Table 1.2 identifies what is considered an imputed value for the variables needed to assign E-sample records to a post-stratum. The values in the table are the HCEF allocation flag values for these variables. The flags can be found on the Person Dual System Estimation E-sample Output Person File.

Table 1.2: Identifying Imputed Values for the E Sample

E-sample Characteristic	Reported Values	Imputed Values
Age	0 = Both Consistent 1 = Age Only 2 = Date of birth only	3 = Inconsistent age and date of birth 4 = Allocated from hot deck 9 = E-sample person did not match to the HCEF
Race	0 = As reported	3= Assigned from race response to Hispanic origin question 4 = Allocated from within household 5 = Allocated from hot deck 9 = E-sample person did not match to the HCEF
Hispanic Origin	0 = 1 reported origin 1 = 2 reported origin 2 = 3 reported origin	3 = Assigned Hispanic Origin from race code 4 = Allocated from within household 5 = Allocated from hot deck (surname used) 6 = Allocated from hot deck (surname not used) 9 = E-sample person did not match to the HCEF
Sex	0 = As reported	1 = From first name 4 = Allocated from hot deck 5 = Allocated from consistency check 9 = E-sample person did not match to the HCEF
Tenure	0 = As reported	1 = Assigned by consistency check 4 = Allocated from hot deck 9 = E-sample person did not match to the HCEF

Race/Hispanic Origin Domain

The Race/Hispanic origin domain assignment is hierarchical. See Haines (2000) for more detail.

Domain 1 (American Indian or Alaska Native on reservations) includes:

- Any person living on a reservation indicating American Indian or Alaska Native either as their single race or as one of many races, regardless of their Hispanic origin.

Domain 2 (American Indian or Alaska Native off reservations) includes:

- Any person living in Indian Country[1] but not on a reservation who indicates American Indian or Alaska Native either as their single race or as one of many races, regardless of their Hispanic origin.

- Any non-Hispanic person not living in Indian Country who indicates American Indian or Alaska Native as their single race.

Domain 3 (Hispanic) includes:

- All Hispanic persons who are not included in Domains 1 or 2.

- All Hispanic persons who self-identify with three or more races (excluding American Indian or Alaska Native in Indian Country).

- All Hispanic persons who do not live in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander, regardless of whether they identify with a single or multiple race.

---

[1] Indian Country is land considered (either wholly or partially) on an American Indian reservation/trust land, Tribal Jurisdiction Statistical Area, Tribal Designated Statistical Area, or Alaska Native Village Statistical Area. For Census 2000, Tribal Jurisdiction Statistical Area has been formally renamed as Oklahoma Tribal Statistical Area.

Domain 4 (Non-Hispanic Black) includes:

- Any non-Hispanic person who indicates Black as their only race.
- Any person identifying with a combination of Black and American Indian or Alaska Native not in Indian Country.
- Any person who indicates Black and another single race group (Native Hawaiian or Pacific Islander, Asian, White, or "Some other race").
- All Non-Hispanic Black persons who do not live in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander.

Domain 5 (Native Hawaiian or Pacific Islander) includes:

- Any non-Hispanic person indicating the single race Native Hawaiian or Pacific Islander.
- Any non-Hispanic person who identifies with the race combination of Native Hawaiian or Pacific Islander and American Indian or Alaska Native not in Indian Country.
- Any non-Hispanic person who identifies with the race combination of Native Hawaiian or Pacific Islander and Asian.
- All persons living in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander, regardless of their Hispanic origin and whether they identify with a single or multiple race.

Domain 6 (Non-Hispanic Asian) includes:

- Any non-Hispanic person indicating Asian as their single race.
- Any person who self-identifies with Asian and American Indian or Alaska Native not in Indian Country.

Domain 7 (Non-Hispanic White or "Some other race") includes:

- Any Non-Hispanic person indicating White or "Some other race" as their single race.
- Any Non-Hispanic person who self-identifies with both American Indian or Alaska Native not in Indian Country and White or "Some other race."
- Any person who self-identifies with Asian and White or Asian and "Some other race."
- Any non-Hispanic person who self-identifies with three or more races (excluding American Indian or Alaska Native in Indian Country).
- Any Non-Hispanic White or Non-Hispanic "Some other race" person who classifies themselves as Native Hawaiian or Pacific Islander but does not live in Hawaii, regardless of whether they identify with other races.